Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Evaluations of Large Language Models a Bibliometric analysis

Sello Prince Sekwatlakwatla[1], Vusumuzi Malele[2]
[1]sek.prince@gmail.com , [2]Vusi.Malele@nwu.ac.za
Unit for Data Science and Computing School of Computer Science and Information Systems
North-West University Vanderbijlpark, South Africa

Article Information

Submitted : 14 Feb 2024
Reviewed: 17 Feb 2024
Accepted : 27 Feb 2024

Keywords

Large language models, artificial intelligence, ethical implications, bibliometric analysis

Abstract

The development of artificial intelligence (AI) and the increased curiosity about how large language models (LLMs) may maximize an organization's opportunities and the ethical implications of LLMs, such as the ability to generate human-like text, give rise to concerns regarding disinformation and fake news. As a result, it is crucial to develop evaluation benchmarks that take into account the social and ethical implications involved. The great challenges of LLMs lack awareness of their own limitations, yet they persist in producing responses to the best of their capabilities. This often results in seemingly plausible but ultimately incorrect answers, posing challenges to the implementation of reliable generative AI in industry. This paper aims to delve into the evaluation metrics of machine-learning models' performance, specifically focusing on LLM. Therefore, bibliometric analysis utilized to explore and analyze various techniques and methods used in evaluating large language models. Additionally, it sheds light on the specific areas of focus when evaluating these models. The results show that natural language processing systems, classification of information, and computational linguistics are some of the techniques used to evaluate large language models. This work paves the way for future investigations employing extensive language models.

A. Introduction

Large language models (LLMs) utilize Artificial intelligence (AI), deep learning, and vast data sets to generate text, translate between languages, and write diverse content types [1]. Large language models (LLMs) have demonstrated exceptional capabilities in various tasks, gaining attention and utilized in numerous downstream applications [2-3]. The challenges of LLMs lack awareness of their own limitations, yet they persist in producing responses to the best of their capabilities [4]. This often results in seemingly plausible but ultimately incorrect answers, posing challenges in the implementation of reliable generative AI for industry [4-5].

The remarkable achievements of LLMs in the domains of law and finance prompt inquiries regarding the reliability of current assessment criteria. The demand for intricate and demanding assignments necessitates the development of fresh datasets and more advanced benchmarks [6].

The ethical implications of LLMs, such as the ability to generate human-like text, give rise to concerns regarding disinformation and fake news [7]. As a result, it is crucial to develop evaluation benchmarks that take into account the social and ethical implications involved.This paper aims to delve into the evaluation metrics of machine-learning models' performance, specifically focusing on large language models (LLMs). Companies and organizations seeking to automate and improve communication and data processing can greatly benefit from these valuable models [7-9]. The large language model's success is largely due to its ability to numerically, represent words in their context [7-8], a significant improvement over previous attempts to automate psychological assessment from language. Large language models (LLMs) are gaining popularity in research and commercial applications due to their ease of use and lack of specialized hardware or software [9].

Maroteau et al, propose a large language method for improving image sequence consistency and contextual fidelity in Hollywood movie scripts, this approach can substantiated through empirical evaluations.LLMs, or advanced computer models, employ deep learning algorithms to carry out a wide range of tasks such as recognition, search, translation, prediction, speech, generative text, and bots [11]. These models utilize sophisticated techniques to process and generate natural language efficiently [10].The potential of student feedback classification enhanced through the successful execution of complex tasks using a large language model. This was achieved by employing unsupervised pre-training and fine-tuning techniques in the study [10-11]. The study proposes a method for evaluating large language models code generation capabilities and incorrect code suggestions from four large language models, categorized into code problems and understanding problems [12]. Code problems involve errors in code, while understanding problems arise when models do not fully understand user requirements.

The study explores the effectiveness of Chatbots (LLMs) in improving L2 vocabulary learning, highlighting their positive impact on receptive and productive knowledge acquisition and incidental learning, particularly in language education. Oduoye et al, caution against the adoption of LLMs in medical writing because of the potential bias present in the model's creation, which could result in adverse consequences downstream [13-14].

Large language models (LLMs) are generative algorithms that utilize extensive text data to generate natural language responses by predicting logical word choices through probabilistic training; therefore, large language models enhance programming error messages, providing novice-friendly enhancements that surpass original messages in interpretability and actionability, benefiting computing educators in challenging student areas [15-16].

The research investigates the application of Large Language Models (LLMs) as an automated instrument for assessing educational environments, encompassing short-answer evaluation methods, entity extraction, and real-time performance evaluation [17].

To achieve this, the paper utilizes bibliometric analysis to explore and analyze various techniques and methods used in evaluating large language models. Additionally, it sheds light on the specific areas of focus when evaluating these models.In spite of this initial introduction, the present document encompasses a section on method, followed by the result and discussion, and conclusion .

B. Research Method

In order to achieve the aim of the research, the following three (3) analytical methods investigated (see Figure 1).

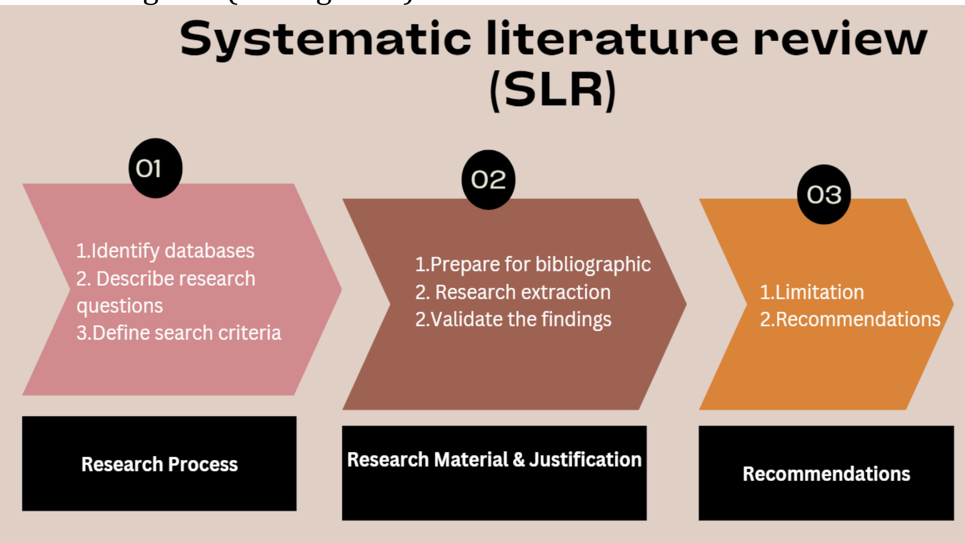
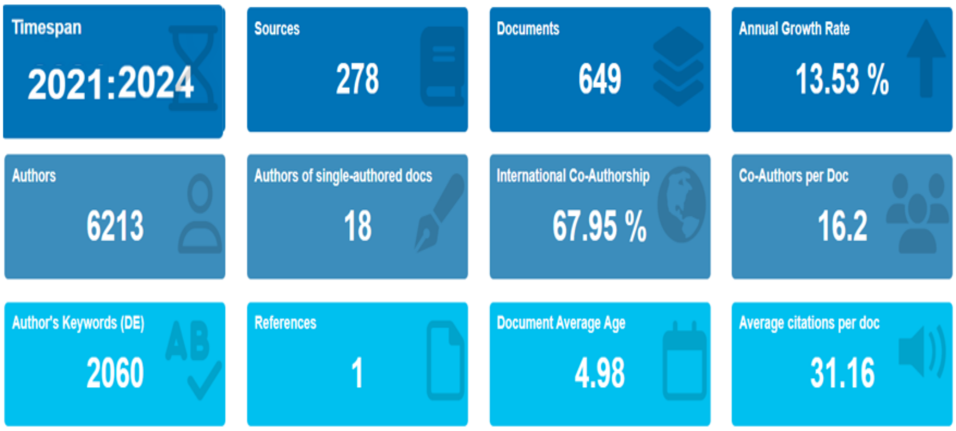


Figure 1. Proposed model

1. Step 1 Research process

The investigation employed various databases such as the Association for Computing Machinerythe (ACM), Web of Service, and Scopus to explore the field of evaluations of large language models. The search conducted using the specific criteria of "Evaluations of large language models." The databases accessed on February 10, 2024. The data collected spanned from January 20, 2021, to January 30, 2024, and encompassed conferences, journals, early-access articles, and magazines that were deemed relevant to the research.



Timespan
2021:2024

Sources
278

Documents
649

Annual Growth Rate
13.53 %

Authors
6213

Authors of single-authored docs
18

International Co-Authorship
67.95 %

Co-Authors per Doc
16.2

Author's Keywords (DE)
2060

References
1

Document Average Age
4.98

Average citations per doc
31.16

Figure 2. Summary of analysis

Only papers that are relevant are listed. With 6213 authors and a 13.53% yearly growth rate, 649 documents with 67.95% international co-authorship were downloaded in this study (see Figure 2).

To guide the research, the following research question utilized, what are techniques for the evaluations of Large Language Models?

2. Step 2 Research Material and justification

The procedure encompasses three main steps: data preparation, data analysis, and generating output. Additionally, the dataset renamed to BibTeX for efficient bibliographic organization. Furthermore, valuable information extracted, and graphs downloaded to aid in decision-making.

3. Step 3 Recommendations

This section identifies the techniques and discusses their limitations based on the findings of the bibliometric analysis.

C. Result and Discussion

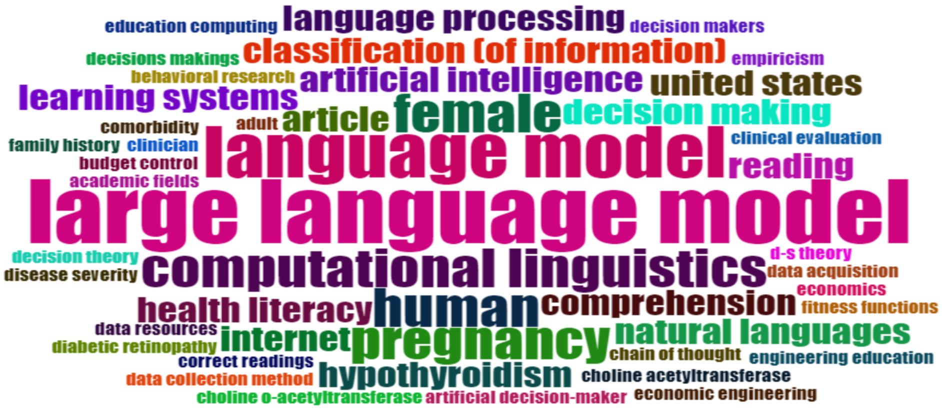


Figure 3. Frequetly search words

This study's bibliometric analysis reveals that future evaluations of large language models may benefit from advancements in computational linguistics, human comprehension, artificial intelligence, decision-making, learning systems, and language processing (see figure 3)

Artificial Intelligence and Decision-Making combines intellectual traditions from computer science and electrical engineering to develop techniques for evaluation of large language models,analyzing and synthesizing systems that interact with the outside world through perception, communication, and action, as well as learning, making decisions, and adapting to changing environments[11].

Many businesses already rely heavily on artificial intelligence, and it is rapidly being utilized to guide global policy and public sector choices [14]. In this regard, the results of this analysis recommend data collection methods,artificial decision makers,fitness functions, and data acquisitions as a tool of evaluation of large language models, linking language models, and learning systems, as well as artificial intelligence, decision-making, learning systems, and language processing.

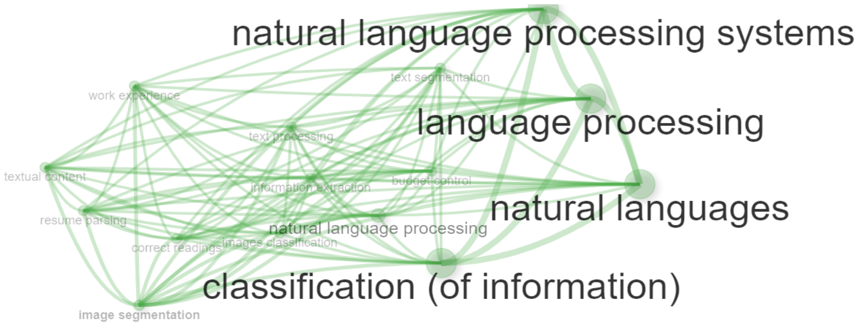


Figure 4. co-network for large language models

Natural language processing systems, language processing, natural languages, and classification of information highlighted in Figure 4 as techniques that may evaluate the data and improve the evaluations of large language models.

Evaluations of large language models can also enable computers and other digital devices to detect, comprehend, and produce text and voice, natural language processing (NLP), combines statistical and machine learning models with computational linguistics, which is rule-based modeling of human language (see figure 4). In this regards Businesses require an effective method for processing the vast volumes of unstructured, text-rich data that they utilize [16]. Until recently, companies were unable to examine the vast majority of natural human language data that was produced online and kept in databases therefore Natural language processing and classification of information is helpful in this situation [17-18].

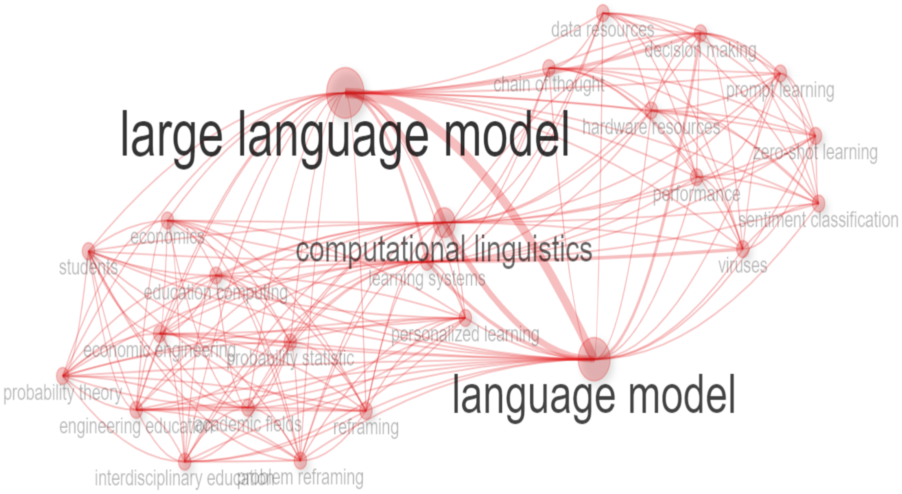


Figure 5. co-network for technique.

Figure 5 illustrates how the link between the big language model and language model processing yields computational linguistics and Learning systems, personalized learning,probability statistics, performance, and reframing are some of the tools that can support the enhancement of evaluations of large language models(see Figure 5), therefore the scientific and technical field of computational linguistics examines spoken and written language from a computer perspective and develops tools for producing and processing language in large quantities or in conversational settings. Given that language is a reflection of the mind, understanding language computationally also illuminates cognitive processes and intelligence for evaluations of large language models [15].

The investigation employed various databases, such as the Association for Computing Machinery (ACM), Web of Service, and Scopus, to explore the field of evaluations of large language models. By using bibliometric analysis, This study's bibliometric analysis reveals that future evaluations of large language models may benefit from advancements in computational linguistics, human comprehension, artificial intelligence, decision-making, learning systems, and language processing. Therefore, to answer the research question, what are the techniques for the evaluation of large language models? The study recommends the combination of the techniques for the enhancement of the results using the following: computational linguistics, human comprehension, artificial intelligence, decision-making, learning systems, and language processing.

D. Conclusion

Large language models (LLMs) have demonstrated exceptional capabilities in various tasks, gaining attention and utilized in numerous downstream applications. The challenges of LLMs lack awareness of their own limitations, yet they persist in producing responses to the best of their capabilities. This often results in seemingly plausible but ultimately incorrect answers, posing challenges in the implementation of reliable generative AI for industry.

This paper aims to delve into the evaluation metrics of machine-learning models' performance, specifically focusing on LLM. Therefore, bibliometric analysis is utilized to explore and analyze various techniques and methods used in evaluating large language models.

The study recommends the combination of the techniques for the enhancement of the results using the following: computational linguistics, human comprehension, artificial intelligence, decision-making, learning systems, and language processing, this study utilized Association for Computing Machinery (ACM), Web of Service, and Scopus, but future expansions may include another research database.

E. Acknowledgment

The article is part of the PhD program in Computer and Information Sciences with Information Technology offered by the School of Computer Science and Information Systems at North-West University, South Africa.

F. References

[1] Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong GU and Si-Qing Chen, An evaluation on large language model outputs: Discourse and memorization, journal of Natural Language Processing Journal,4,2023,[Online].Available: https://doi.org/10.1016/j.nlp.2023.100024

[2] Mannam et al, Large Language Model-Based Neurosurgical Evaluation Matrix: A Novel Scoring Criteria to Assess the Efficacy of ChatGPT as an Educational Tool for Neurosurgery Board Preparation,journal of World Neurosurgery 180 ,2023,e765-e773, [Online].Available: https://doi.org/10.1016/j.wneu.2023.10.043

[3] Zixuan Wang,Paul Denny,Juho Leinonen,Andrew Luxton-Reilly,Leveraging Large Language Models for Analysis of Student Course Feedback,COMPUTE: Proceedings of the 16th Annual ACM India Compute Conference 23,2023,76–79, [Online]. Available: https://doi.org/10.1145/3627217.3627221

[4] H. Su, J. Ai, D. Yu and H. Zhang, An Evaluation Method for Large Language Models' Code Generation Capability,10th International Conference on Dependable Systems and Their Applications (DSA), Tokyo, Japan, 10,2023,831-838, doi: 10.1109/DSA59317.2023.00118.

[5] Zhihui Zhang and Xiaomeng Huang, The impact of chatbots based on large language models on second language vocabulary acquisition 'journal of Heliyon,10,2024,[Online].Available: https://doi.org/10.1016/j.heliyon.2024.e25370

[6] Maciej pankiewicza and Ryan s. bakerb, Large Language Models (GPT) for automating feedback on programming assignments, Proceedings of the st International Conference on Computers in Education. Asia-Pacific Society for ComputersEducation,31,2023,[Online].Available: https://www.researchgate.net/publication/375091243

[7] Olatunde Oduoye M, et al. Algorithmic Bias and research integrity; the role of nonhuman authors in shaping scientific knowledge with respect to artificial intelligence (AI); a perspective. Int J Surg, 109,2023,:2987–90. Doi: 10.1097/JS9.0000000000000552

[8] Haruka Kumagai, Ryosuke Yamaki and Hiroki Naganuma: Story-to-Images Translation: Leveraging Diffusion Models and Large Language Models for Sequence Image Generation, NarSUM: Proceedings of the 2nd Workshop on User-centric Narrative Summarization of Long VideosOctober 23,2023, 57–63 [Online]. Available: https://doi.org/10.1145/3607540.3617144

[9] Leinonen et al,Using Large Language Models to Enhance Programming Error Messages.SIGCSE,Proceedings of the ACM Technical Symposium on Computer Science Education V. 54,2023,563–569. [Online].Available: https://doi.org/10.1145/3545945.3569770

[10] Oscar Kjell, Katarina Kjell and Andrew Schwartz, Beyond rating scales: With targeted evaluation, large language models arepoised for psychological assessment, journal of Psychiatry Research, 333,2024, [Online].Available: https://doi.org/10.1016/j.psychres.2023.115667

[11] Maroteau et al,Evaluation of the impact of large language learning models on articles submitted to Orthopaedics & Traumatology: Surgery & Research (OTSR): A significant increase in the use of artificial intelligence in 2023, journal of Orthopaedics & Traumatology: Surgery & Research,109,2023, [Online].Available: https://doi.org/10.1016/j.otsr.2023.103720

[12] Wynter et al, An evaluation on large language model outputs: Discourse and memorization, journal of Natural Language Processing Journal, 4,2023, [Online].Available: https://doi.org/10.1016/j.nlp.2023.100024

[13] Rana AlShaikh,Norah Al-Malki and Maida Almasre,The implementation of the cognitive theory of multimedia learning in the design and evaluation of an AI educational video assistant utilizing large language models, journal of Heliyon, 10 (2024). [Online].Available: https://doi.org/10.1016/j.heliyon.2024.e25361

[14] Chenggang Mi and Shaoliang Xie, Language relatedness evaluation for multilingual neural machine translation, journal of Neurocomputing, 570 ,2024, [Online].Available: https://doi.org/10.1016/j.neucom.2023.127115

[15] Paul Brie,Nicolas Burny,Arthur Sluÿters and Jean Vanderdonckt,Evaluating a Large Language Model on Searching for GUI Layouts,Proceedings of the ACM on Human-Computer Interaction,71,2023, 1–37 [Online].Available: https://doi.org/10.1145/3593230

[16] Paul Brie,Nicolas Burny,Arthur Sluÿters and Jean Vanderdonckt,Evaluating a Large Language Model on Searching for GUI Layouts,Proceedings of the ACM on Human-Computer Interaction,71,2023, 1–37 [Online].Available: https://doi.org/10.1145/3593230

[17] Panagiotis Tsoutsanis and Aristotelis Tsoutsanis, Evaluation of Large language model performance on the Multi-Specialty Recruitment Assessment (MSRA) exam, journal of Computers in Biology and Medicine, 168,2024, [Online].Available: https://doi.org/10.1016/j.compbiomed.2023.107794

[18] Sandeep Kumar and Arun Solanki,Natural Language Processing System using CWS Pipeline for Extraction of Linguistic Features Procedia Computer

Science,218,2023,1768-1777,[Online].Available: https://doi.org/10.1016/j.procs.2023.01.155